DOCUMENT RESUME

ED 420 700 TM 028 375

AUTHOR Evans, Thomas J.
TITLE Standard Setting Models for High School Graduation Competency Tests.
PUB DATE 1998-01-00
NOTE 13p.; Paper presented at the Learning '98 National Conference for Bilingual Educators (1st, San Pedro Sula, Honduras, January 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; *Graduation Requirements; *High School Students; High Schools; *Minimum Competency Testing; *Standards; *Test Use; Testing Programs
IDENTIFIERS *Standard Setting

ABSTRACT

This paper explores common concerns about competency testing as they relate to the certification of high school graduates seeking a diploma in the United States. Competency testing is widespread in the United States, with 40 states engaged in competency testing in at least one grade. In general, and particularly for graduation requirements, the certification of minimum competency is the objective, as fears that the minimum levels defined would become the accepted standards for all students have been discredited. A number of standard setting methods exist to determine standards for minimum competency. Numerous test-centered continuum models have been proposed for competency testing programs, and the most common of these are reviewed. Two examinee-centered continuum models are also described. Several authors have compared standard setting methods, as it is apparent that the standard setting procedures used to arrive at justifiable standards for competency tests vary in method and results. Careful consideration should be given to the choice of any single standard setting method, and the wisest course of action may be to use several procedures to attempt to reach convergence at an appropriate cut score. (Contains 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 700

STANDARD SETTING MODELS FOR HIGH SCHOOL GRADUATION COMPETENCY TESTS

by

Thomas J. Evans, Ed.D.
Evans Consulting Group

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Thomas J. Evans
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper Prepared For

LEARNING '98:
The First National Conference
For Bilingual Educators In Honduras

San Pedro Sula
Honduras

January 1998

TM028375

INTRODUCTION

Competency tests exist for the general purpose of ensuring that individuals are sufficiently qualified in specific academic areas. Students in many states are required to take a graduation examination before receiving a high school diploma. A passing score represents to society that the successful examinee is certifiably knowledgeable to a predetermined level. However, a number of issues and questions surround the practice of competency testing such as: (a) What level of knowledge should a certified student possess?; (b) Why set standards at all since they are arbitrary in nature?; (c) What are the methods by which the passing score dividing mastery from failure are determined?; and, (d) Is one standard setting method "better" than another? This paper will investigate these common concerns surrounding competency testing as they relate to the certification of high school graduates seeking a diploma in the United States.

STUDENT COMPETENCY TESTING IN THE UNITED STATES

Competency testing of students is a widespread practice throughout the United States. Pipho (cited in Linn, 1989) reported that there were 40 states engaged in competency testing in at least one grade. In total, such examinations represented the testing of student competency across the entire grade span of Kindergarten through 12th grade. Though student competency testing has existed for hundreds of years throughout world history in various forms to serve multiple purposes, the phenomena currently observed in the United States has originated from a surge of interest by legislators and laypersons during the 1970s into apparent shortfalls of public education. The expected benefits from competency testing include: "(1) restore confidence in the high

school diploma, (2) involve the public in education, (3) improve teaching and learning, (4) serve a diagnostic, remedial function, and (5) provide a mechanism of accountability" (Gorth and Perkins 1979, p. 12).

Clearly the focus of student graduation competency-testing is on assessing the demonstration of a requisite minimum amount of knowledge before certification is granted. Airasian, Pedulla, and Madaus (cited in Linn, 1989, p. 486) described competency testing for United States students as "a certification mechanism whereby a pupil must demonstrate that he/she has mastered certain minimal skills in order to receive a high school diploma". However, the interest in certifying *minimum competence* rather than some form of academic excellence led critics to argue that minimum standards would replace maximum standards thereby endangering standards for all students. Despite this argument, the testing of minimum competency has prevailed and, hence, necessitated devising methods to determine defensible minimum standards.

STANDARD SETTING PHILOSOPHY

A number of standard setting methods exist that vary both procedurally and in the final standard produced. Disparate techniques produce different standards since the nature of the standard setting process is, in essence, a judgmental activity. Jaeger (1976) describes this activity as follows:

> All standard-setting is judgmental. No amount of data collection, data analysis and model building can replace the ultimate judgmental act of deciding which performances are meritorious or acceptable and which are unacceptable or inadequate. All that varies is the proximity of the judgment-determining data to the original performance (p. 2).

Such judgmental decision making between meritorious or unacceptable performances elicited diverse responses from experts in educational measurement. For example, Glass (1978) and Burton (1978) believed this judgmental act so sufficiently arbitrary in nature as to preclude the use of any derived standards. Hambleton (1978, 1980), Popham (1978), Scriven (1978), and Shepard (1976, 1979) offered pragmatic arguments for the necessity of setting standards. They believed that standards could serve to aide educational decision making notwithstanding the unresolved philosophical and methodological problems inherent to standard setting procedures. Considering theoretical concerns in the light of practical consequences, Mehrens (1987) acknowledged that while standards delineating mastery/non-mastery are arbitrary in nature, these and other dichotomous decisions of mastery must be made in practical life situations. Mehrens explained that:

> (1) Although mastery is a continuous, not dichotomous, construct, we are forced to make dichotomous decisions... We do need to decide who knows enough to graduate from high school. Even if everyone graduates, there has still been a categorical decision as long as the philosophical or practical possibility of failure exists. If one can conceptualize performance so poor that the performer should not graduate, then theoretically a cutoff score exists. (2) Although setting a cutting score may be arbitrary, it need not be capricious. Setting a cutting score on tests is usually less capricious a choice than many other categorical decisions that are made in life (pp. 126-127).

CLASSIFICATION OF STANDARD SETTING METHODS

Meskauskas (1976) proposed the classification of standard setting methods into "state models" and "continuum models". State models assume that an examinee either possesses some degree of the competence or completely lacks any competence. State models have not been used

to an appreciable extent compared to continuum models. Continuum models assume that the construct measured is a continuous variable that can take on any value over a given numerical interval.

Jaeger (cited in Linn, 1989) suggested further dividing continuum models in "test-centered models" and "examinee-centered models". The distinction between these two categories of models rests in the entity about which expert judgments are made. Specifically, test-centered models require judgments about the content of the tests, considering the test holistically or the items separately, whereas examinee-centered models require judgments about the competence of the examinees with respect to the competencies of interest.

STANDARD SETTING METHODS

Test-Centered Continuum Models

Numerous test-centered continuum models have been proposed for use in competency-testing programs though several are used with greater regularity. The first of these is the *Angoff procedure*. Angoff (cited in Linn 1989) proposed that a panel of expert judges separately examine each item on a competency test and estimate:

> ... the probability that the 'minimally acceptable' person would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score (p. 493).

A second popular standard setting method is *Ebel's procedure* (Crocker and Algina, 1986). This method also relies on expert judgments, in this case, regarding what percentage of a certain category of test items a minimally competent person could be expected to answer correctly. The categories are established by filling cells of a "difficulty" (usually with three levels) by "relevance" (usually with four levels) test item grid. The resulting standard is:

> ... a weighted average of the proportions recommended by the judges for each category of items. That is, the proportion recommended for each cell is multiplied by the number or items in that cell, and the products are them summed. This sum of products is divided by the total number of items on the test, to produce a weighted average percentage (Jaeger cited in Linn 1989, p. 494).

When multiple judges are involved, the final cut score can be determined by calculating a mean weighted percentage for the entire group of the judges.

A third standard setting procedure used in competency-testing is *Jaeger's procedure*. Expert judges are asked separately to determine if every examinee should be able to answer the particular test item under consideration. A judge's recommended standard is the number of items that he or she believed every examinee should know. The test standard for a sample of judges is the median of the standards recommended by the judges in that group. Since several samples of judges are selected, the operational standard is determined as the lowest of the median recommended standards for all samples of judges.

A final test-centered continuum model that receives wide spread use is the *Nedelsky procedure*. This procedure requires judges to conceptualize a minimally competent examinee and predict which response options such a person should be able to eliminate as incorrect for each multiple choice item on a test. A minimum pass level is computed for each item and is equal to

the reciprocal of the remaining number of response options. Ultimately, a judge's recommended standard is the sum of all minimum pass levels for each test item. An average of the recommended standards for a sample of judges is used as the test standard.

Examinee-Centered Continuum Models

There are two examinee-centered continuum models proposed by Zieky and Livingston (1977) that are used often in competency testing. They are the borderline-group procedure and contrasting-groups procedure.

With the *borderline-group procedure*, judges (for example, teachers) familiar with the competence of students are asked to classify them into three categories: (1) competent; (2) borderline; and (3) incompetent. The test is then administered to the students and the test standard determined as the median score for the borderline examinee group.

The *contrasting-groups method* also uses judges to identify groups of competent and incompetent students prior to the test administration. The actual test data is used to determine the test standard which can be done in a number of ways. Hambleton and Eignor (1980) suggested determining the test standard as the point of intersection between the competent students' frequency distribution with that of the incompetent students' frequency distribution.

COMPARISON OF STANDARD SETTING METHODS

Standard setting procedures may produce varying test standards even when applied to the

same test. In one case, Mills (1985 cited in Crocker and Algina, 1986) used the Angoff procedure, contrasting groups, and borderline groups method to compare standards obtained by using test item judgments (as opposed to judgments of examinee's performance). This study highlighted:

> ...that when three or more methods are used, it may be possible to obtain some convergence between at least two of the methods. For example, standards from Angoff's method and the contrasting groups method were more consistent with each other than with the standards from the borderline group method for most of the cases reported (p. 417).

Hambleton (1980), Koffler (1980), and Shepard (1980; 1984) suggested using several standard setting methods for any particular study, and by considering the results obtained -- along with any relevant extra-statistical factors -- the appropriate standard should be set.

As the preceding references pointed out, it is imperative to have guidelines to assist in the critical decision of which standard setting procedure to use since different methods will yield different cut scores. To this end, Berk (1986) produced a useful consumer's guide to setting performance standards on criterion-referenced tests.

Berk (1986) provided a brief description of twenty-three continuum standard setting methods, including their advantages and disadvantages. He subclassified them into eleven judgmental, seven judgmental-empirical, and five empirical-judgmental standard-setting methods. A consumer's guide style of evaluation followed for each method and was composed of a total of ten technical adequacy and practicability criteria.

Berk (1986) defined *technical adequacy* as "the extent to which a method satisfies certain psychometric and statistical standards that would render it defensible to experts on standard setting" (p. 140). Berk presented six technical criteria to use to evaluate a particular standard

setting method which included: (1) yield appropriate information; (2) be sensitive to examinee performance; (3) be sensitive to instruction or training; (4) be statistically sound; (5) identify the true standard; and (6) yield decision validity evidence.

Berk (1986) defined *practicability* as "the ease with which a standard-setting method can be implemented, computed, and interpreted" (p. 141). The four practicability criteria included the assessment of the extent to which the method was: (1) easy to implement; (2) easy to compute; (3) easy to interpret to laypeople; and (4) credible to laypeople.

Berk (1986) evaluated the twenty-three standard setting methods across all ten criteria with mean values calculated individually and together for technical adequacy criteria and practicability criteria. The methods were compared overall, and those rated highest were identified. Berk commented, "the Angoff method appears to offer the best balance between technical adequacy and practicability" (p. 170). The contrasting-groups method was rated the highest among all methods for technical adequacy. The informed judgement method obtained the highest rating overall.

Berk (1986) noted that the stakes are higher for decisions of mastery/nonmastery when considering high school graduation, than decisions within the context of a classroom. Yet, diverse conditions preclude a recommendation of a best method for each certification test. Berk recommended a five point "eclectic judgmental-empirical method" that contained "some form of judgmental analysis", "conceptual and computational simplicity", and the best elements of the other methods presented. Berk concluded:

> If performance data are not available, the Angoff method is recommended. If iterations are unfeasible, the informed judgement method should be considered... Despite the technical attractiveness of the contrasting-groups method, it must be relegated to a lesser position in the rankings due to the

political realities of the standard-setting enterprise (p. 172).

SUMMARY

Competency testing of students is an established procedure throughout many states. The fear that minimum levels of competence would become the accepted maximum level for all students has been discredited over time. Nevertheless, the standard setting procedures used to arrive at justifiable standards for competency tests vary in method and results. Careful consideration should be given to the choice of any single standard setting procedure, and perhaps the wisest course of action would be to use several procedures in an attempt to reach convergence at an appropriate cut score.

REFERENCES

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Burton, N. (1978). Societal standards. Journal of Educational Measurement, 15, 263-271.

Crocker, L., & Algina, J. (1986). Introduction to Classical & modern test theory. New York: Holt, Rinehart and Winston.

Glass, G. (1978). Standards and criteria. Journal of Educational Measurement, 15, 277-290.

Gorth, W.P., & Perkins, M.R. (1979). A study of minimum competency testing programs (Final program development resource document). Amherst, MA: National Evaluation Systems.

Hambleton, R.K. (1980). Test score validity and standard setting methods. In R. S. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore Johns Hopkins University Press.

Hambleton, R.K., & Eignor, D.R. (1980). Competency test development, validation, and standard setting. In R.M. Jaeger & C.K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396). Berkley, CA: McCutchan.

Hambleton, R.K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 18, 22-27.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Linn, R. L. (1989). Educational measurement. National Council on Measurement in Education and American Council on Education. 3rd ed. New York: American Council on Education.

Mehrens, W. A. and Lehmann, I. J. (1987). Using standardized tests in education. (3rd). New York: Holt, Rinehart and Winston.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 45, 133-158.

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 273-275.

Shepard, L. A. (1976). Setting standards and living with them. Florida Journal of Educational Research, 18, 23-32.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore: Johns Hopkins University Press.

Shepard, L. A. (1980). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of research in education: Vol. 8 (pp. 30-82). Washington, DC: American Educational Research Association.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda and J. R. Sanders (Eds.). Practices and problems in competency-based measurement. National Council of Measurement in Education.

Zieky, M. J., & Livingston, S. A. (1977). Manual for setting standards on the basic skills assessment tests. Princeton, NJ: Educational Testing Service.